IS 6733
Deep Learning on Cloud Platforms

Lecture 1a

Big Data Overview

Dr. Yuanxiong Guo

Department of Information Systems and Cyber Security

Acknowledgement

- Some slides are adapted from Professor Widom's Instructional Odyssey
    - www.professorwidom.org

- Thanks to Prof. Jennifer Widom for sharing the slides

What Does “Big Data” Mean?

(1) Collecting large amounts of data

Via computers, sensors, people, events

(2) Doing something with it

Making decisions, confirming hypotheses, gaining insights, predicting future …

“Data Science” = Going from (1) to (2)

Big Data is Here to Stay

- Ability to collect data will only increase
- Ability to analyze data will only improve

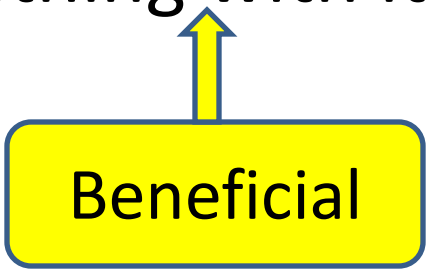
This Overview

- Promises of Big Data
  - Applications and services
- Big Data tools and techniques
  - Database management systems
  - Data mining and machine learning
- Pitfalls of Big Data
  - Correlation and causation
  - Underfitting and overfitting
  - Privacy and a few others
- Big Data systems and platforms

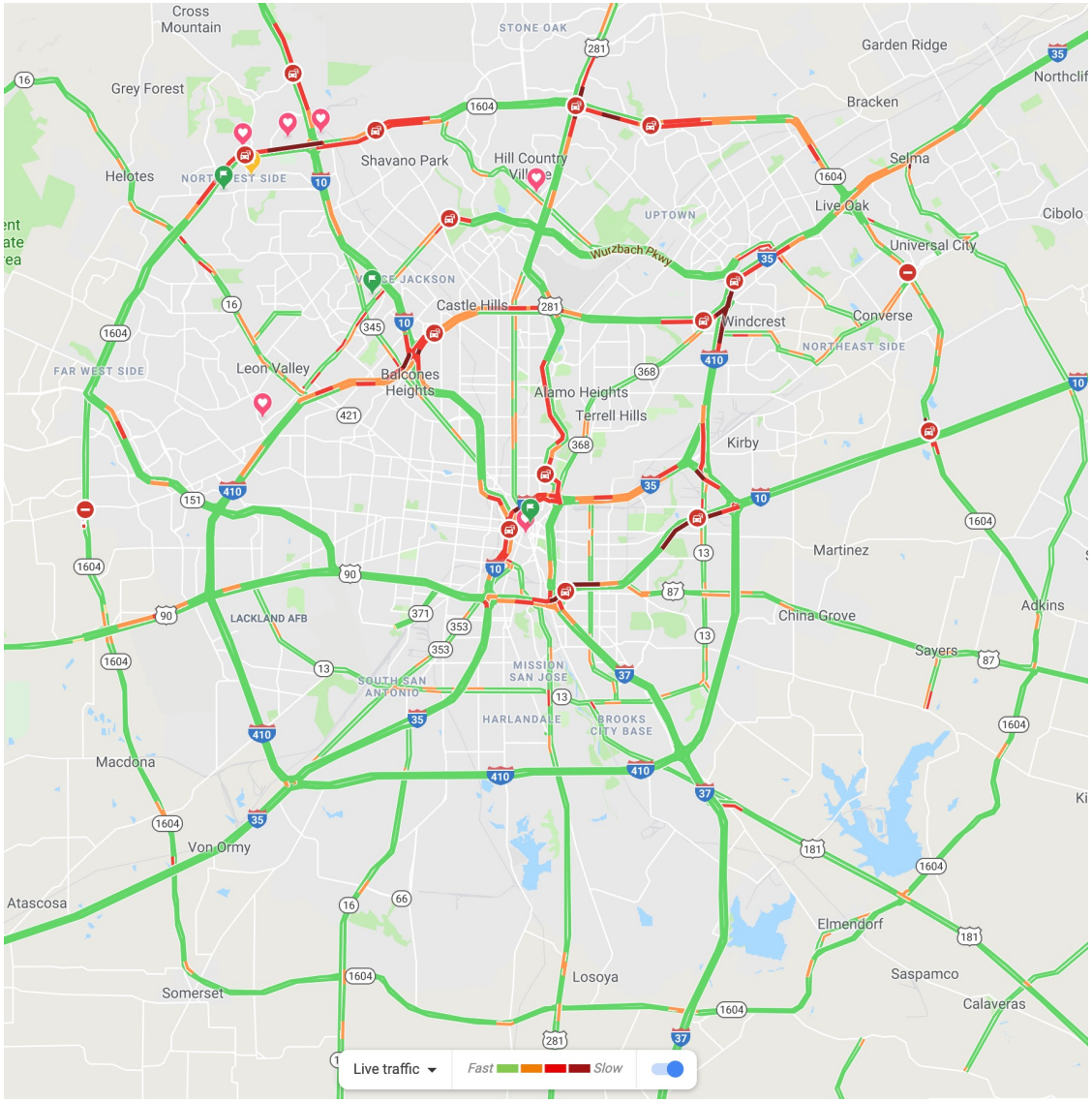
Promises of Big Data

(1) Collecting large amounts of data

(2) Doing something with it



Traffic

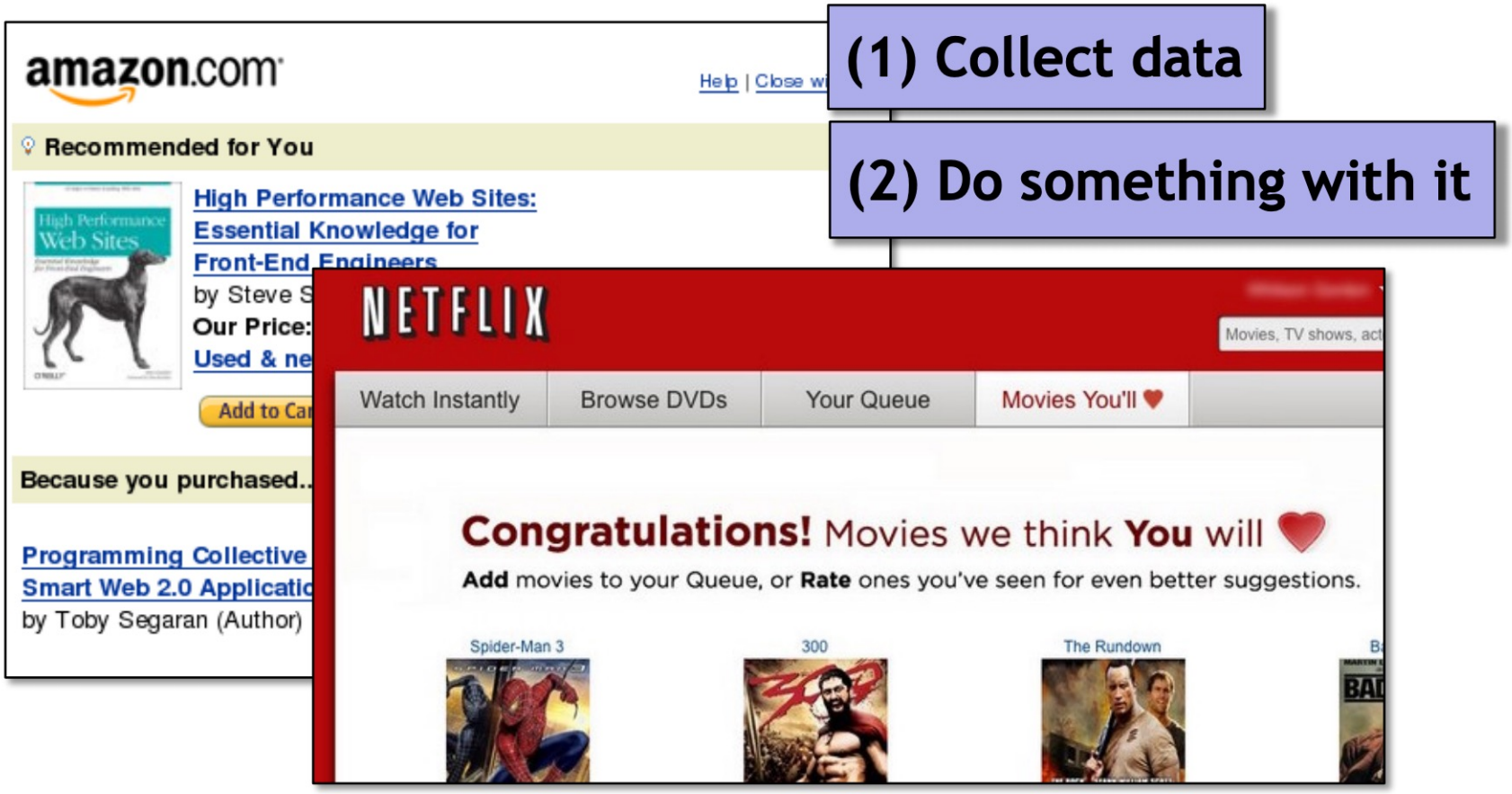


(1) Collect data

(2) Do something with it

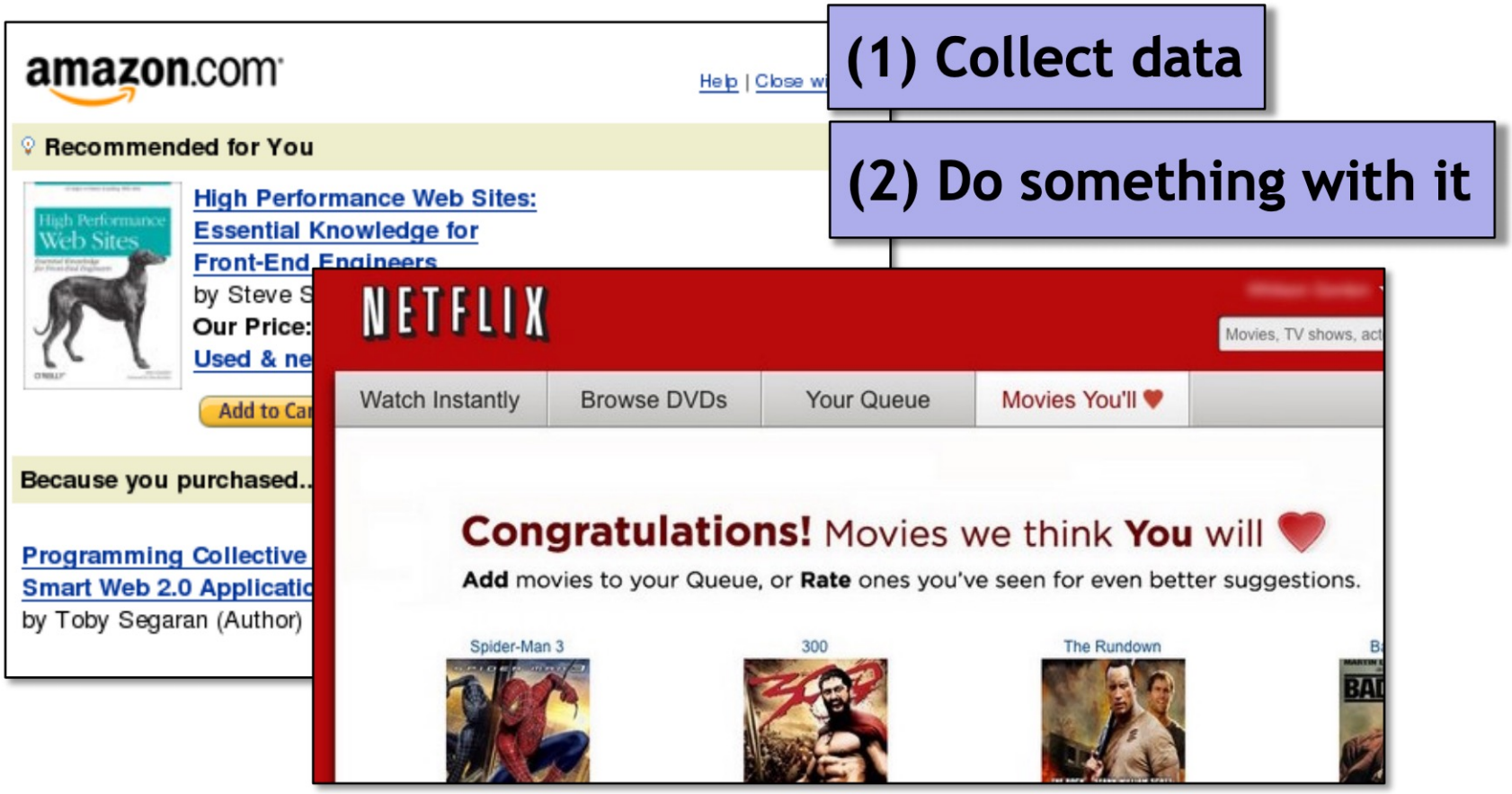
Friday, 08/23/2019, 5 pm CST

Recommender Systems



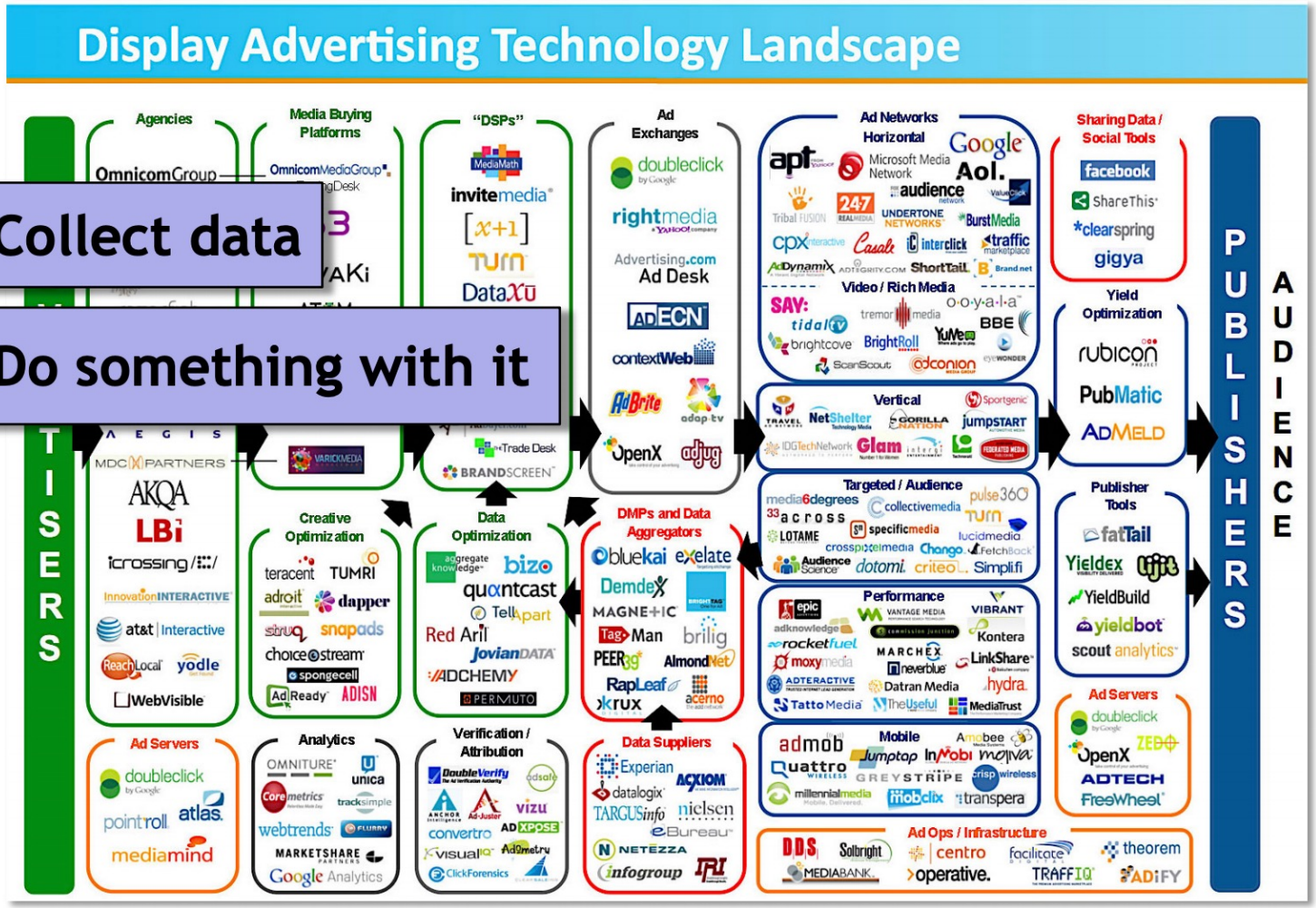
+ music, news, friends, romantic partners, and many more!

Recommender Systems

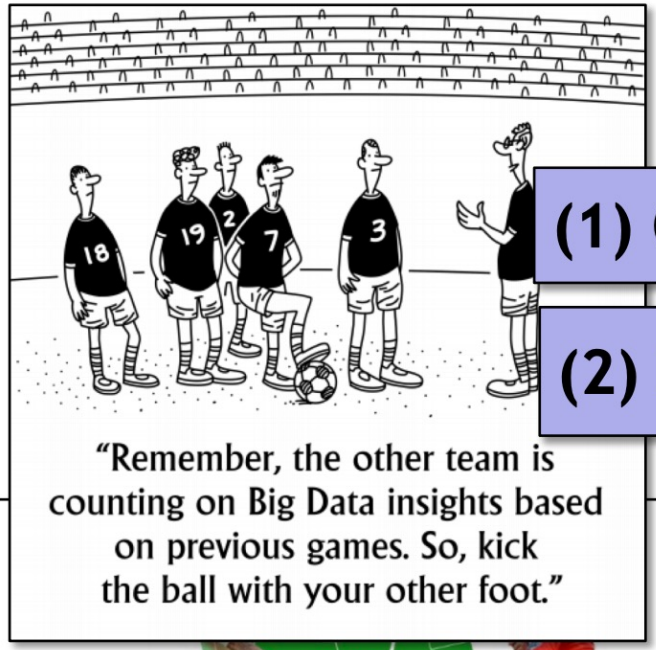


+ music, news, friends, romantic partners, and many more!

Online Advertising

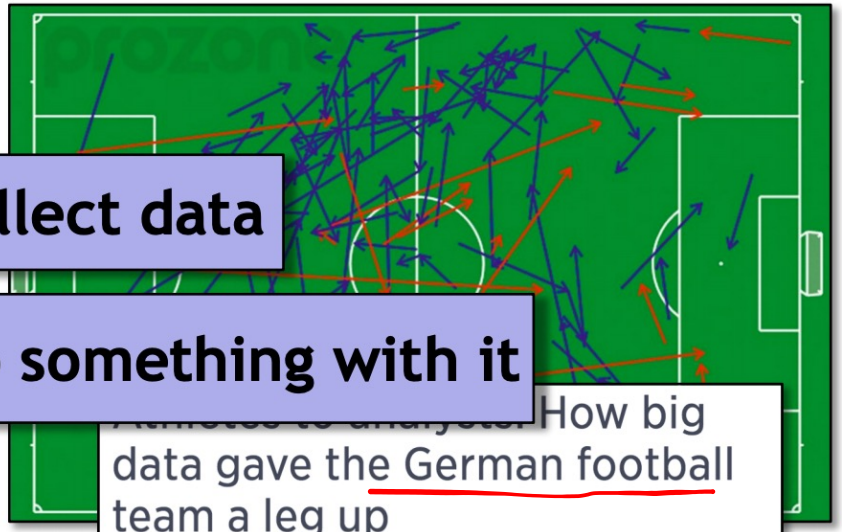


Sports



(1) Collect data

(2) Do something with it

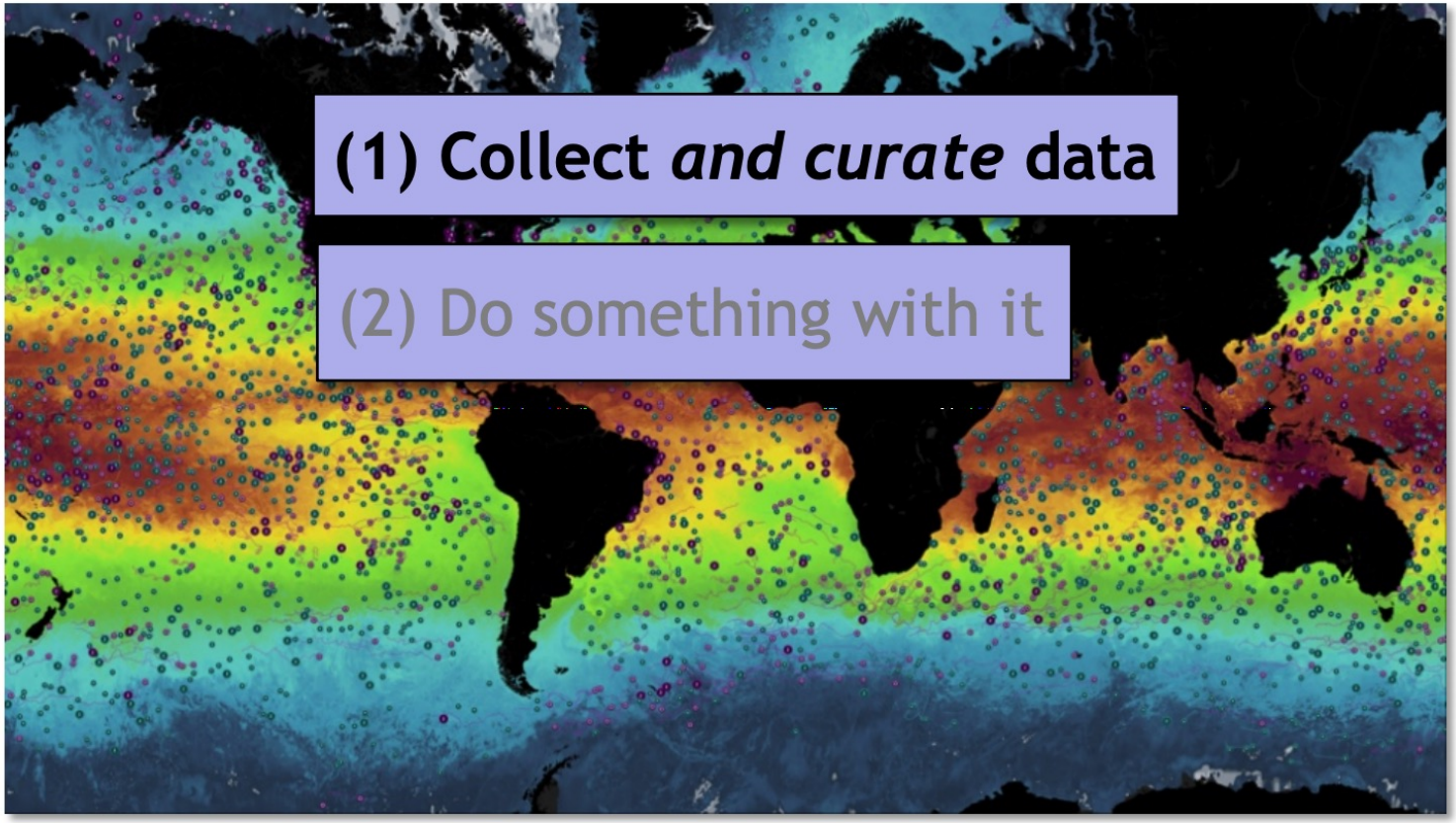


How big data gave the German football team a leg up

Saheli Roy Choudhury | @sahelirc
Thursday, 7 Jul 2016 | 12:39 AM ET

CNBC

Ocean Health

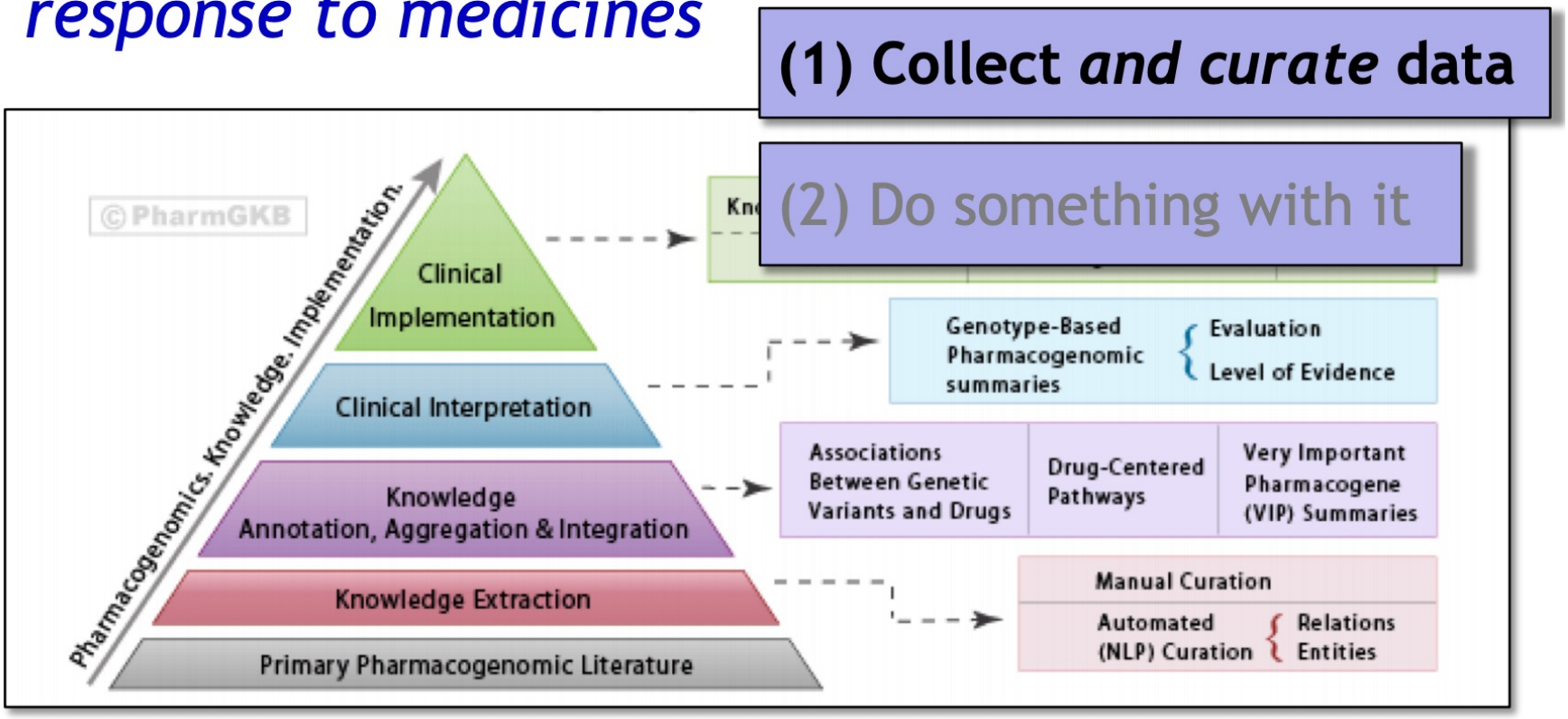


44,000 sensors, over 2 billion measurements

Physical, chemical, biological …

Genetics-Medicine Relationship

PharmGKB collects, curates, and disseminates knowledge about how human genetics affects response to medicines



And Many More

- Weather prediction
- Medical diagnosis
- Financial markets
- Resource management
- Computational social science
- Smart buildings and cities
- …

The list goes on and one, and it's still early days

Big Data Tools and Techniques

- Basic Data Manipulation and Analysis
  - Performing well-defined computations or asking well-defined questions (“queries”)
- Data Mining
  - Looking for patterns in data
- Machine Learning
  - Using data to build models and make predictions
- Data Visualization
  - Graphical depiction of data
- Data Collection and Preparation

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Average January low temperature for each country over last 20 years
- Number of items over $100 bought by females between ages 20 and 30
- Frequency of specific medicine relieving specific symptoms
- The ten stocks whose price varied the most over the past year

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Ave… …h cou…
- Nur… fen…
- Fre… specific symptoms
- The ten stocks whose price varied the most over the past year

- Spreadsheets
- Relational (SQL) database systems
- "NoSQL" / scalable systems
- Programming languages with big-data support (e.g., Python, R)

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People who like movie X also like movie Y
- Patients who respond well to medicines X and Y also respond well to medicine Z
- Students going to the same university are frequently online friends
- Wealthier people are moving from cities to suburbs

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People ... ie Y
- Patien ... es X and Y
- Studer ... are frequently online friends
- Wealthier people are moving from cities to suburbs

- Frequent item-sets
- Association rules
- Specialized techniques for networks, text, multimedia

Machine Learning

Using data to build models and make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on the SAT
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Machine Learning

Using data to build models and make predictions

- Customers w… over age 20 are likely to res… rtisement
- Students wi… re predicted to do well on the…

- Regression
- Classification
- Clustering

- Roughly: Basic data analysis and data mining give answers from the available data, while machine learning uses the available data to make predictions about missing or future data

Data Visualization

"A picture is worth a thousand words"

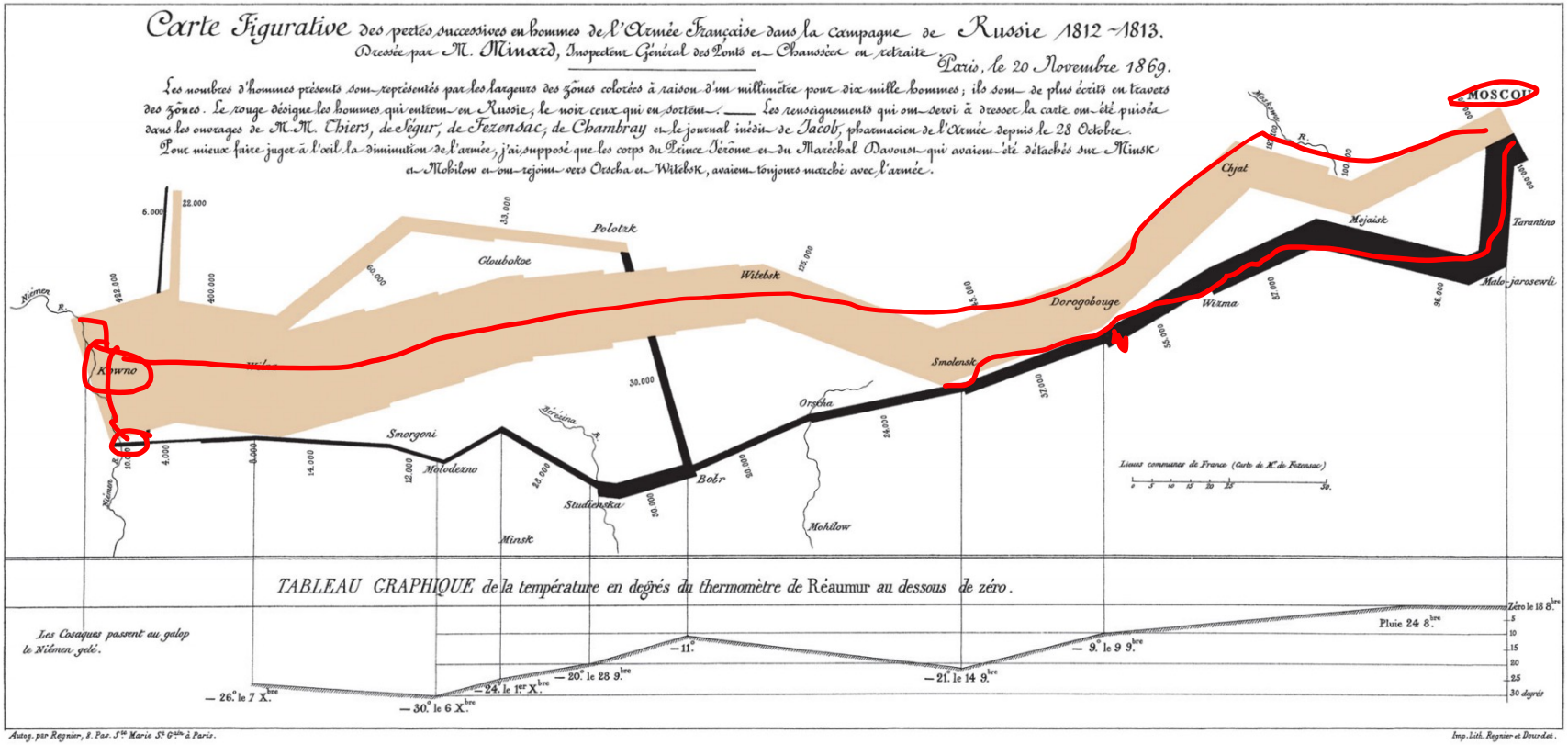
Data Visualization

"A picture is worth a ~~thousand words~~"

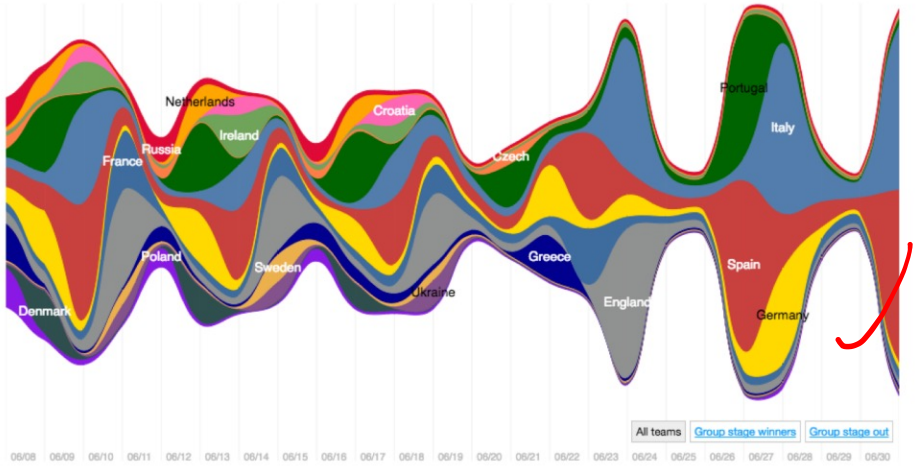
trillion data points

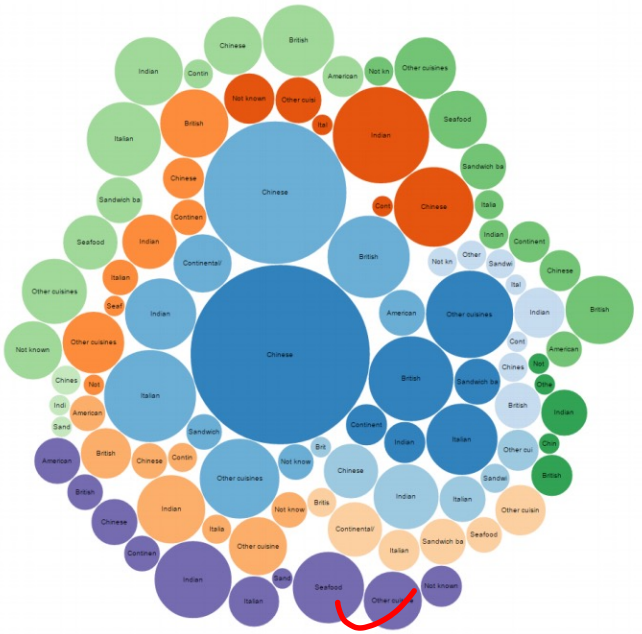
Early Data Visualization

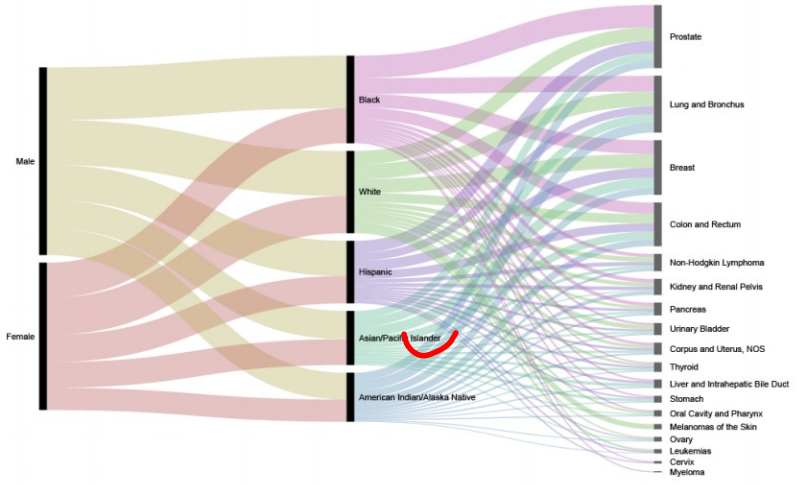
Napoleon's Army (Year 1812)



Fancy Data Visualization



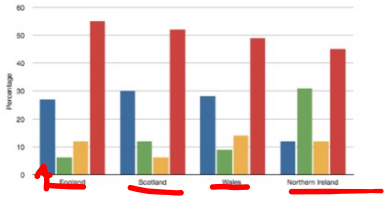


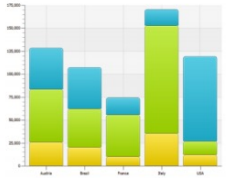


Basic Data Visualization

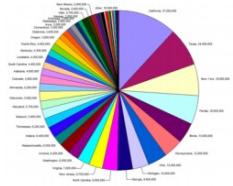
Don’t underestimate the power of basic visualizations

- Bar charts

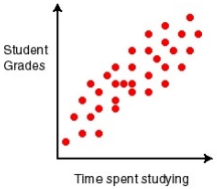


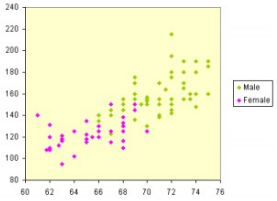


- Pie charts

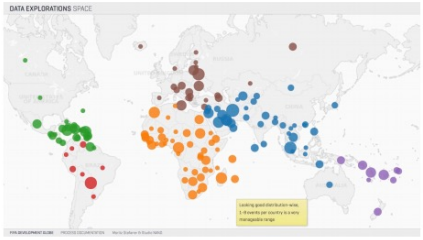


- Scatterplots

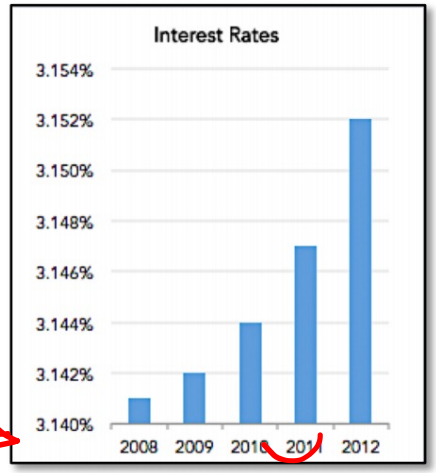


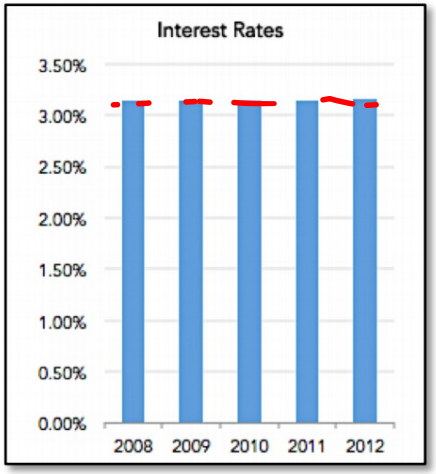


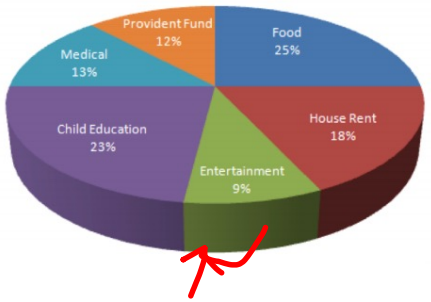
- Maps

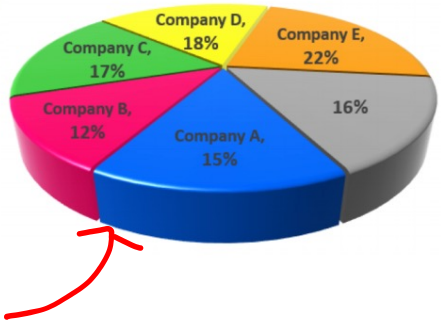


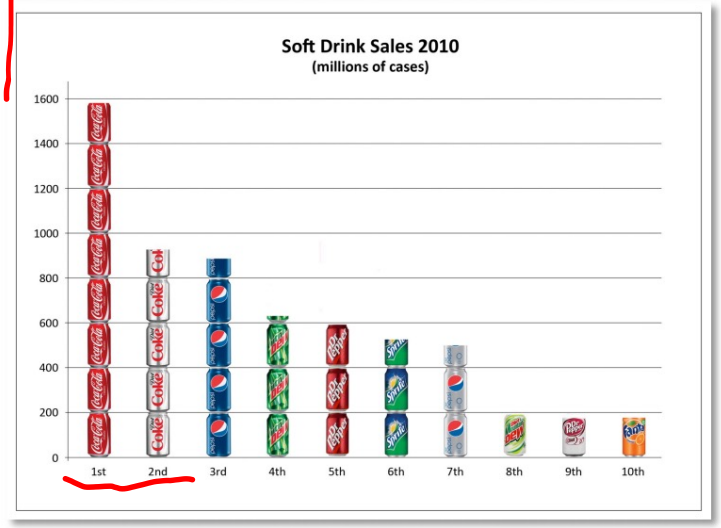
Misleading Data Visualization











Data Collection and Preparation

The "dirty" secret of Big Data

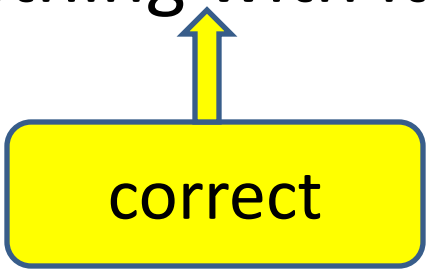
- Extracting data from difficult sources
- Filling in missing values
- Removing suspicious data
- Making formats, encoding, and units consistent
- De-duplicating and matching

Data preparation often consumes 80% or more of the effort in a Big Data project

Pitfalls of Big Data

(1) Collect large amounts of data

(2) Doing something with it



Correlation and Causation

- Data analysis, data mining, and machine learning can reveal relationships between data values

- Correlation – Values track each other
  - Height and Shoe Size
  - Grades and SAT Scores
- Causation – One value directly influences another
  - Education Level → Starting Salary
  - Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

- Correlation – Values track each other
  - Height and Shoe Size
  - Grades and SAT Scores
- Causation – One value directly influences another
  - Education Level → Starting Salary
  - Temperature → Cold Drink Sales

Correlation and Causation

"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B
  - Homeless population and crime rate
    - Confounding variable: unemployment
  - Forgetfulness and poor eyesight
    - Confounding variable: age
  - Height and shoe size
  - Grades and SAT scores

Correlation and Causation

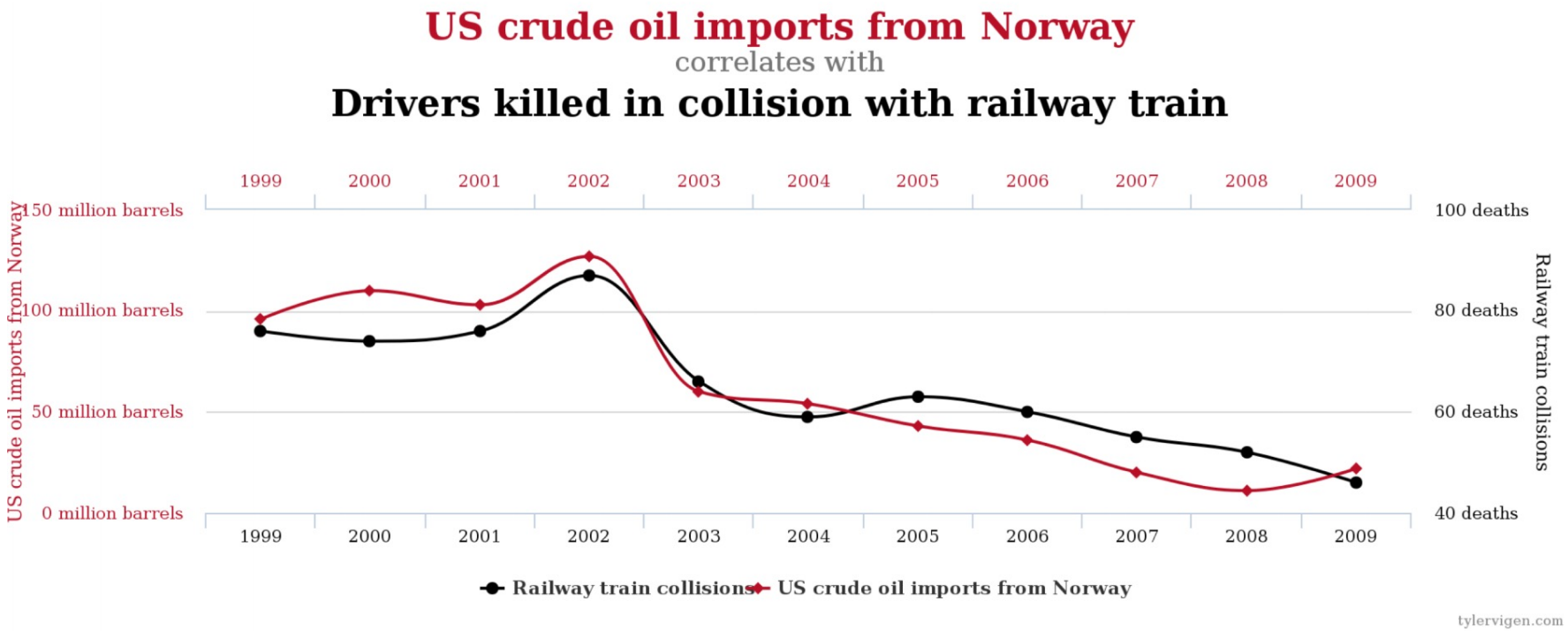
"Correlation does not imply causation"

- Correlation can be result of causation from a hidden "confounding variable"
- A and B are correlated because there's a hidden C such that C → A and C → B

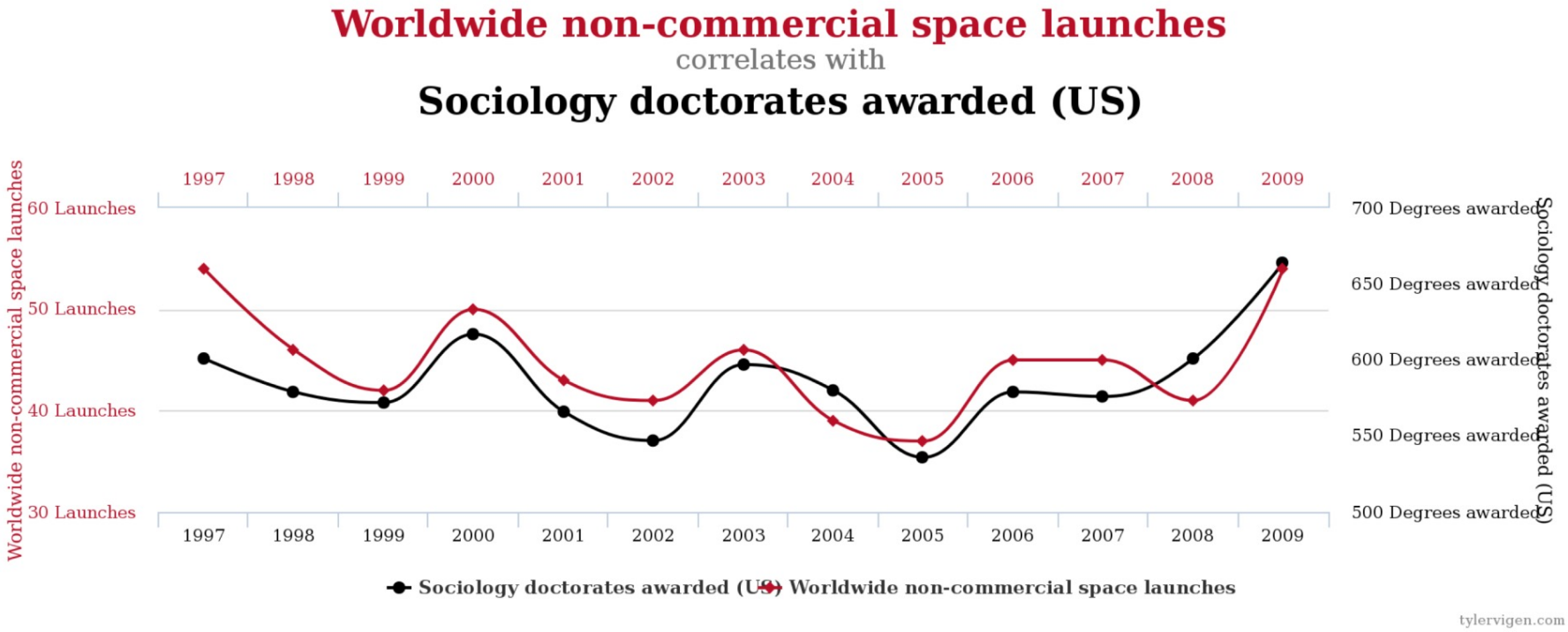
- Correlation is usually "easy" to test
- Causation is impossible to test

Correlation and Causation

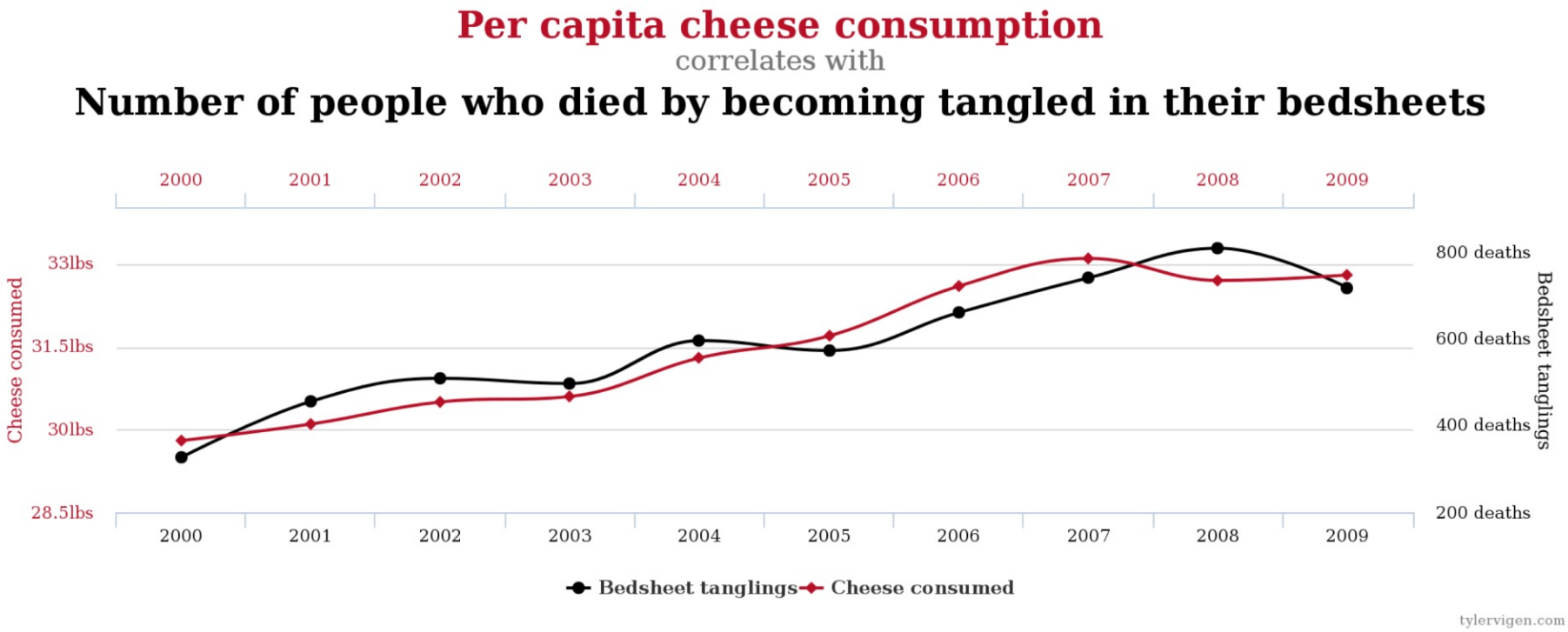
Surprising Correlation #1



Surprising Correlation #2



Surprising Correlation #3



“Surprising Correlation” Website

http://www.tylervigen.com

Machine Learning

Machine learning uses data to create a "model" and uses model to make predictions

- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on the SAT
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

Underfitting

Model used for prediction is **too simplistic**

- 60% of men and 70% of women responded to an advertisement, therefore all future ads should go to women
- If a furniture item has four legs and flat top, it is a dinning room table
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains

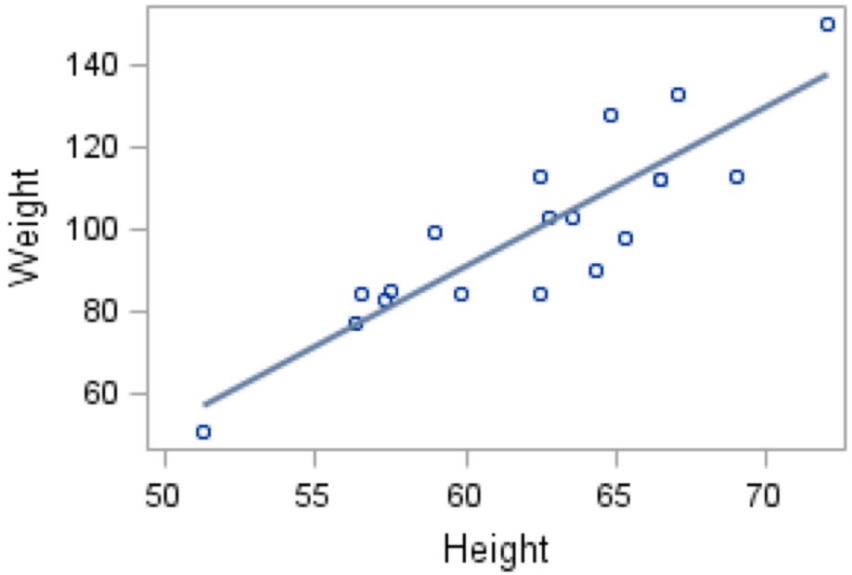
Overfitting

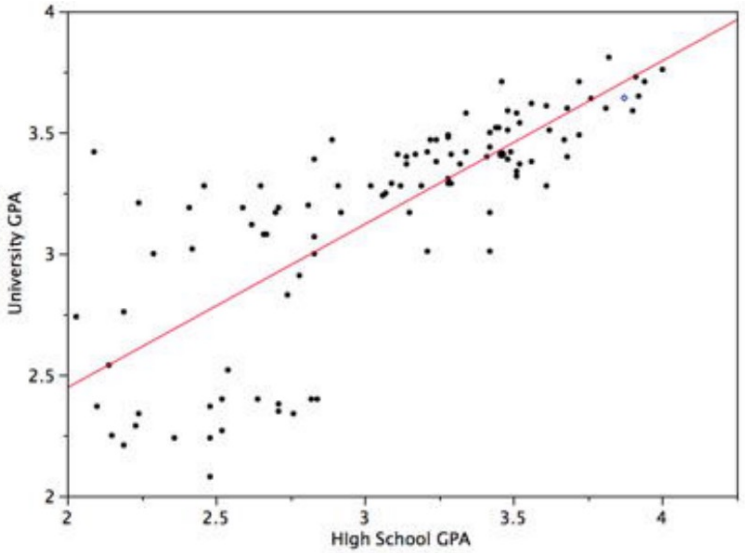
Model used for predictions is **too specific**

- The best targets for an advertisement are married women between 25 and 27 years with short black hair, one child, and one pet dog
- If a furniture item has four 100 cm legs with decoration and a flat polished wooden top with rounded edges then it is a dining room table

Regression

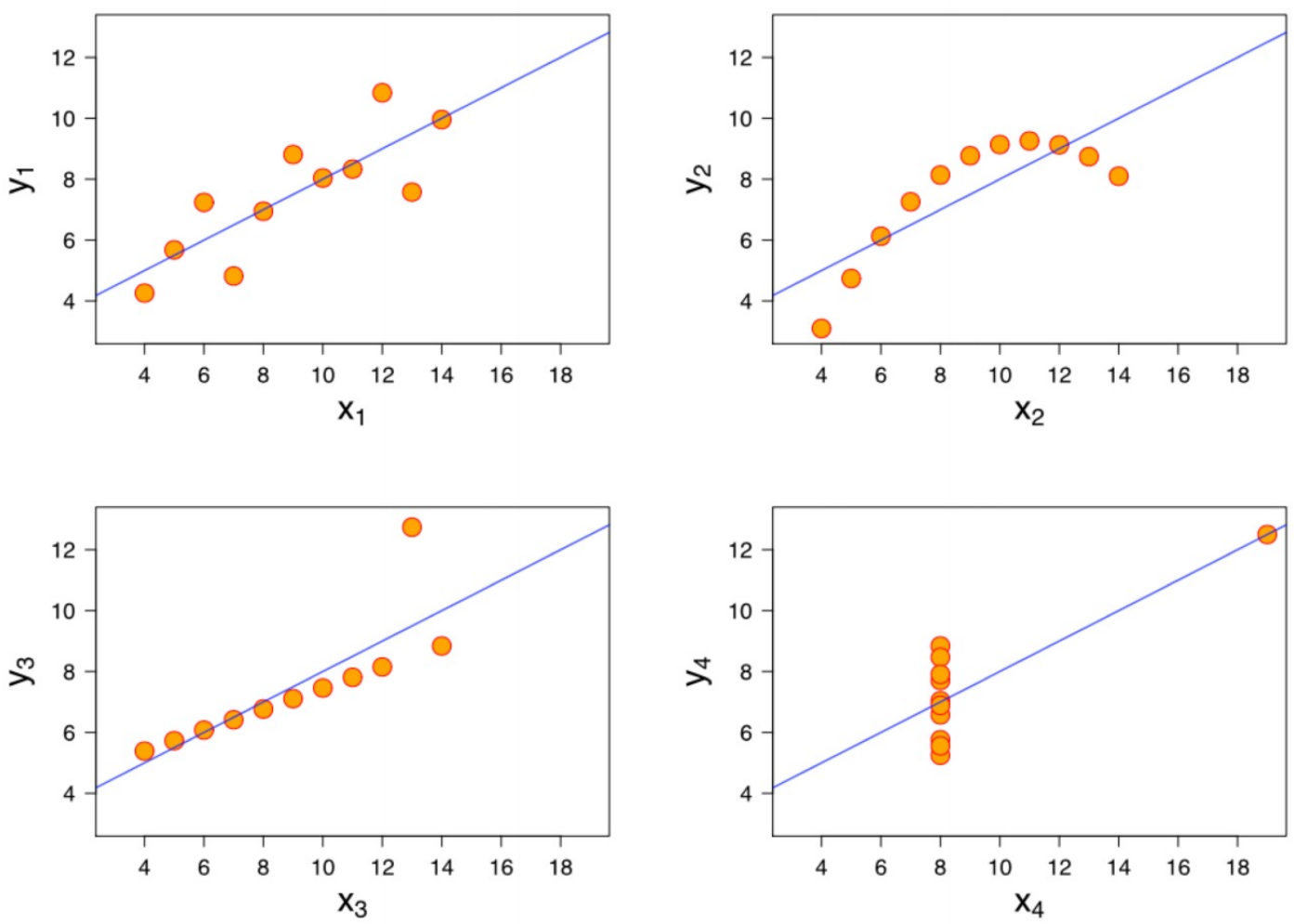
- Fit a line or curve to a set of points (model)
- Use model to predict values for new points





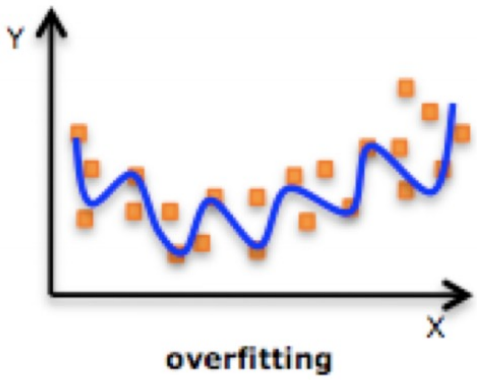
Underfitting

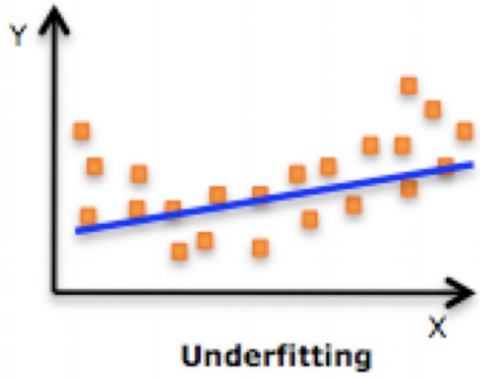
Model is **too simplistic**

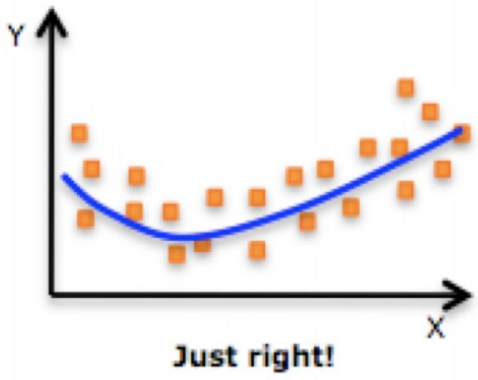


Overfitting

Model is **too specific**







Big Data Scam: Soccer Match Prediction

- Friday: receive email from “Psychic Sally” predicting which teams will be the winners in the weekend’s five soccer matches. She’s right about all of them!
- Same thing the following weekend: five games, all winners predicted correctly
- And the following one: five more correct
- Fourth Friday: Sally offers to give you her predictions for the coming weekend’s games, for a fee

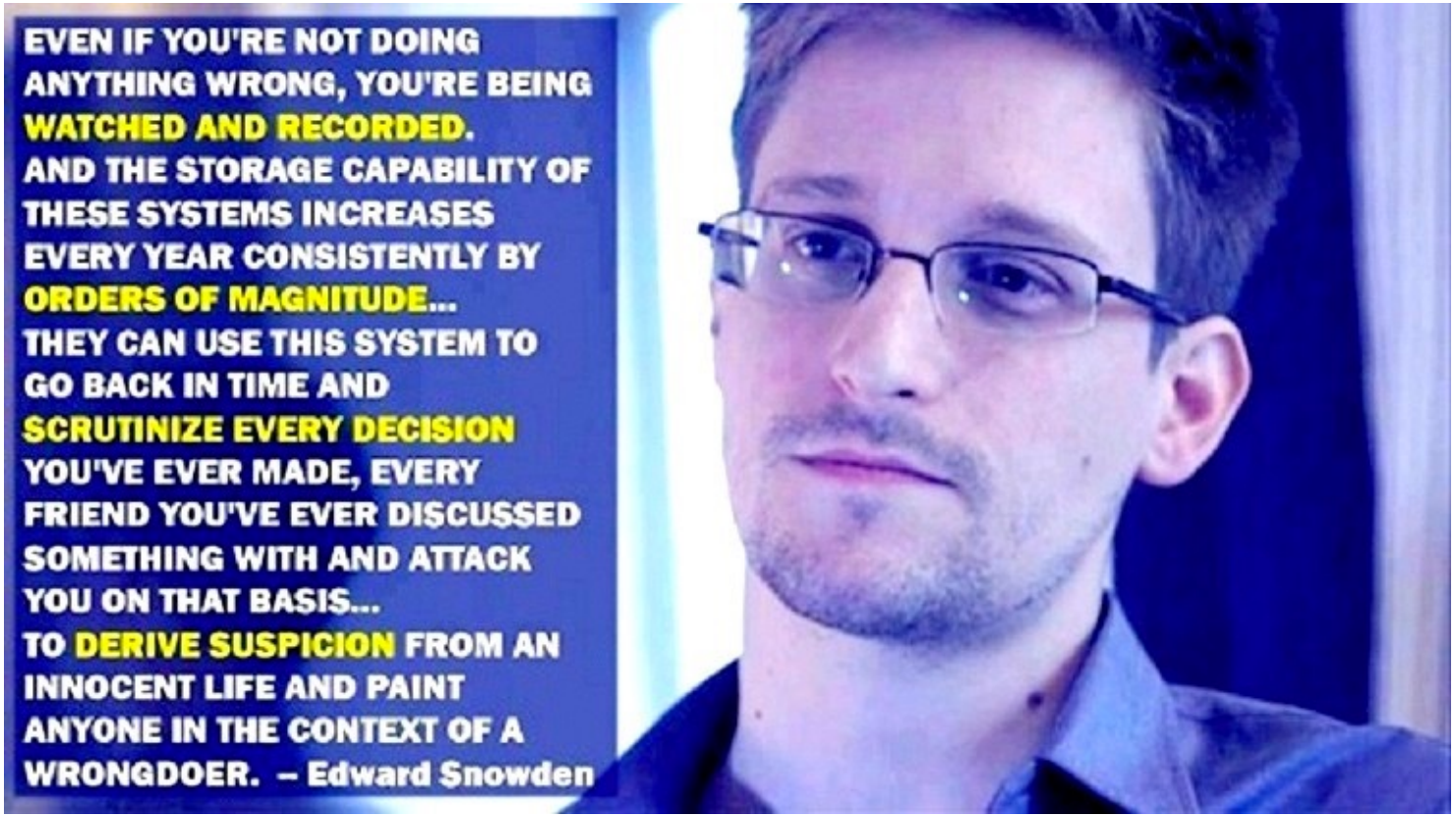
Should you do it?

Big Data Scam: Soccer Match Prediction

How many contacts must Sally start with on week one to ensure she has 100 potential buyers by week four, i.e., 100 people who received 15 correct predicted winners?

Data Privacy

- Individual data collected covertly
  - Edward Snowden, “metadata” argument



Data Privacy

- Individual data collected legally but used questionably
  - Individual "digital footprints" are enormous
  - Target stores pregnancy mailing
  - Facebook–Cambridge Analytica data scandal

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill

Tech

Welcome to The Not-So Private Parts where technology & privacy collide

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target **TGT +0%**TGT +0%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

TARGET

Facebook

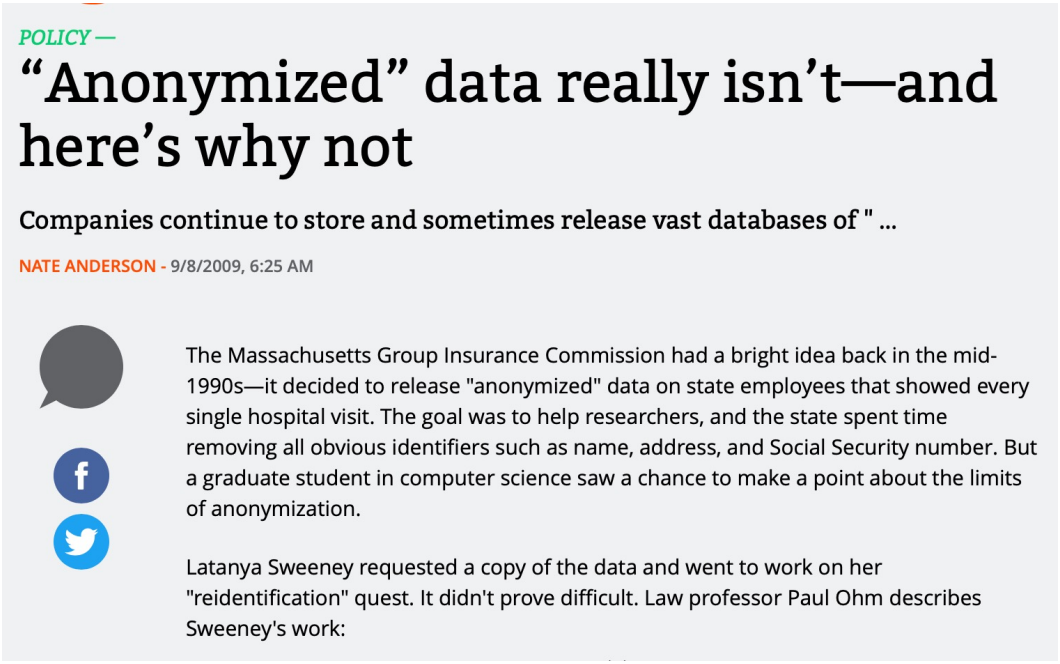
The Cambridge Analytica scandal changed the world – but it didn't change Facebook

A year after devastating revelations of data misuse, Mark Zuckerberg still hasn't fulfilled his promises to reform

▲ Mark Zuckerberg testifies before on Capitol Hill in 2018 following the privacy scandal. Photograph: Xinhua / Barcroft Images

Data Privacy

- Individual data deduced from “anonymous” public data
  - Governor of Massachusetts William Weld health record
  - 5-digit zip code, birth date, gender can uniquely identify 87 percent of the population in US



POLICY —

“Anonymized” data really isn’t—and here’s why not

Companies continue to store and sometimes release vast databases of " ...

NATE ANDERSON - 9/8/2009, 6:25 AM

The Massachusetts Group Insurance Commission had a bright idea back in the mid-1990s—it decided to release "anonymized" data on state employees that showed every single hospital visit. The goal was to help researchers, and the state spent time removing all obvious identifiers such as name, address, and Social Security number. But a graduate student in computer science saw a chance to make a point about the limits of anonymization.

Latanya Sweeney requested a copy of the data and went to work on her "reidentification" quest. It didn't prove difficult. Law professor Paul Ohm describes Sweeney's work:

Languages, Systems, Platforms

- Spreadsheets
  - Surprisingly versatile and powerful for data analysis tasks, but not truly big data
- Programming languages with big-data support
  - R Language: powerful statistical features
  - Python: general-purpose language with R-like add-ons (Pandas, SciPy, scikit-learn)

Languages, Systems, Platforms

- Relational Database Management Systems
  - Also called RDBMS, SQL Systems
  - Long-standing solution for reliability, efficiency, powerful query processing
  - Works for all but truly extreme data sizes, or highly unstructured data
- "NoSQL" Systems
  - Distributed/scalable processing, unstructured data
  - Key-value row stores (e.g., Cassandra, Dynamo)
  - Document databases (e.g., MongoDB, CouchDB)
  - Graph databases (e.g., Neo4J, Giraph)

Languages, Systems, Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark generalized data flow
- Systems for data preparation
- Systems for data visualization

Languages, Systems, Platforms

- Data processing in the cloud
  - Amazon Web Services, Google Cloud, Microsoft Azure
  - Data storage
  - Data processing: SQL, Hadoop, Spark
  - Machine learning libraries
  - Integration with visualization systems

Big Data: How Big is Big?

- Complete works for William Shakespeare
  - 5 megabytes
- Average individual
  - 50 gigabytes (10,000 Shakespeares)
- USA Library of Congress
  - 10 terabytes (2 million Shakespeares)
- Uploaded to Facebook daily
  - 1 petabyte (200 million Shakespeares)
- Produced by humanity daily
  - 2.5 exabytes (500 trillion Shakespeares)

Size Isn’t Everything

- Tools and techniques apply to data of all sizes
- Big insights can come from small/medium data

Some applications actually do need twenty Spark servers in the cloud. More often a laptop with SQL, Python, or simple spreadsheets does the job.

Big Data Overview

Questions?